DOCUMENT RESUME

ED 344 937 TM 018 299

AUTHOR Ackerman, Terry A.; Evans, John A.
TITLE An Investigation of the Relationship between Reliability, Power, and the Type I Error Rate of the Mantel-Haenszel and Simultaneous Item Bias Detection Procedures.
PUB DATE Apr 92
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Equations (Mathematics); *Error of Measurement; *Item Bias; *Mathematical Models; Monte Carlo Methods; Raw Scores; *Sample Size; Test Items; *Test Reliability
IDENTIFIERS Ability Estimates; *Mantel Haenszel Procedure; Power (Statistics); *Simultaneous Item Bias Procedure; Type I Errors

ABSTRACT

The relationship between levels of reliability and the power of two bias and differential item functioning (DIF) detection methods is examined. Both methods, the Mantel-Haenszel (MH) procedure of P. W. Holland and D. T. Thayer (1988) and the Simultaneous Item Bias (SIB) procedure of R. Shealy and W. Stout (1991), use examinees' raw scores as a conditioning variable in the computation of differential performance between two groups of interest. As a result, the extent to which examinees' observed scores accurately reflect their true abilities plays an important role. If examinees are misrepresented by their observed score (as for a test with low reliability) then the ability of bias detection methods to determine item bias may not be very accurate. Results of Monte Carlo studies (40-item test, 720 testing conditions) suggest that for a fixed-length test, the power of both statistics increases moderately as reliability is increased and substantially as sample size is increased. However, the combination of small sample sizes and high reliability results in a decrease of power. For most of the simulated conditions, the MH and SIB procedures have very similar rates of correctly rejecting the biased item. Sixteen plots illustrate the discussion. There is a 15-item list of references. (SLD)

An Investigation of the Relationship Between Reliability, Power and the Type I Error Rate of the Mantel-Haenszel and Simultaneous Item Bias Detection Procedures

Terry A. Ackerman
John A. Evans
University of Illinois

Paper presented at the 1992 NCME Annual Meeting, San Francisco, CA., April 21, 1992.

Abstract

This study examines the relationship between levels of reliability and the power of two bias and differential item functioning (DIF) detection methods. Both methods, the Mantel-Haenszel (MH) (Holland & Thayer, 1988) and the Simultaneous Item Bias (SIB) (Shealy & Stout, 1991), use examinees' raw scores as a conditioning variable in the computation of differential performance between two groups of interest. As a result, the extent to which examinees' observed scores accurately reflect their true abilities plays an important role. If examinees are misrepresented by their observed score (as for a test with low reliability) then the ability of bias detection methods to determine item bias may not be very accurate. Results suggest that for a fixed length test, the power of both statistics increases moderately as reliability is increased and substantially sample size increased. However, the combination of small sample sizes and high relibility resulted in a decrease of power. For most of the simulated conditions the MH procedure and SIB had very similar rates of correctly rejecting the biased item.

An Investigation of the Relationship Between Reliability, Power and the Type I Error Rate of the Mantel-Haenszel and Simultaneous Item Bias Detection Procedures

Objectives of the Study

The purpose of this study was to provide the testing practitioner with information concerning the interaction between test reliability and the accuracy of two item bias and differential item functioning (DIF) detection procedures.[1] The power and Type I error rate of two bias detection methods, the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988) and the Simultaneous Item Bias (SIB) detection procedure (Shealy & Stout, 1991), were examined. The MH procedure has developed into a nonparametric benchmark test that is widely used by many testing practitioners. SIB, a relatively new procedure, is also nonparametric. Unlike the MH procedure, SIB can evaluate the collective bias of more than one item.

Both procedures use raw score as a conditioning variable to form groups of comparable ability examinees. Consequently, it is imperative that the test be equally reliable for both groups and that each examinee's observed score be an accurate indication of their true ability. The concern that prompted this study was that if the reliability of a test were to decrease, the power of MH and/or SIB to detect bias may suffer. The shape of the observed score distribution is a function of the test reliability (cf. Lord, 1953). If a test has low reliability (e.g., $\rho_{xx'}$ =.70) the resulting observed score distribution tends to be leptokurtic with a relatively small variance. As the items become more discriminating and reliability increases the observed score distribution changes to a more platykurtic, uniform shape. At very high levels of reliability ($\rho_{xx'}$ = .95), the distribution becomes U-shaped with examinees being grouped in the tails. Hence, the level of reliability, because of its effect on how examinees are spread out along the raw score scale, could effect the power of the two procedures to accurately detect biased items.

A second area of concern when using these procedures is the number of examinees in the two groups of interest. Often practitioners do not have the luxury of having a large sample of minority subjects, who are usually the focal group of interest. It is quite common for these

groups to be greatly under-represented in the group total examinee population. Thus, the effects of different ratios of focal versus reference group sizes were studied.

A third issue that was addressed in this study was the amount of bias. Because this study is basically focusing on the power of each test as reliability and sample sizes are varied, the effect size or amount of bias had to be varied. In this study the effect size was defined as the amount of angular difference between the measurement direction (i.e., direction of maximum information) of the item and the measurement direction of the test. This concept will be discussed in detail later.

The final factor that was varied in this study was the number of biased items. In cases of no bias the Type I error rates of both procedures were examined. When one item was simulated as being biased the power of each procedure to make a correct rejection was investigated. For multiple biased items the study focused only on the power of SIB, considering its performance as the reliability and the sample sizes of reference and focal groups were varied in selected combinations.

Theoretical Background

Multidimensional IRT

For simplicity, in this paper bias will be examined from a two-dimensional perspective in which one dimension represents the pure, intended-to-be-measured ability, denoted by θ and the other dimension represents the nuisance abilities, denoted by η. The ability η represents a skill that is not intended to be measured, but may be used by examinees to solve an item with a potential for bias. The work of Reckase (1986), which formally defines multidimensional item response theory (MIRT) item characteristics, provides an excellent foundation from which to examine the interaction between multidimensional items and the underlying multidimensional ability distributions for groups of interest.

Reckase's work is based upon the MIRT compensatory model (M2PL) which for the purposes of this paper will be expressed in terms of the true ability dimension, θ, and the nuisance dimension, η. The probability of a correct response to item *i* by examinee *j* can be written as

$$P(X_{ij}=1 \mid a_i, d_i, \theta_j, \eta_j) = \frac{e^{(a_{1i}\theta_j + a_{2i}\eta_j + d_i)}}{1.0 + e^{(a_{1i}\theta_j + a_{2i}\eta_j + d_i)}} \quad (1)$$

where X_{ij} is the score (0,1) on item *i* by person *j*, a_i is the vector of item discrimination parameters, d_i is a scalar difficulty parameter of item *i*, and θ_j, η_j is the vector of ability parameters for person *j*.

In a two-dimensional latent ability space (e.g., math and verbal ability dimensions), the a_{1i} and a_{2i} vectors designate the composite of θ and η that item *i* is measuring. If $a_{1i} = a_{2i}$, both dimensions would be measured equally well. However, if $a_{1i} = 0$ and $a_{2i} = 1.0$, discrimination would occur only along the η dimension. If all of the items in a test are measuring exactly the same (θ, η) composite (i.e., the same "direction" in the (θ, η) coordinate system), the test would be strictly unidimensional. The more varied the composites that are being assessed, the more multidimensional the test.

Reckase's (1986) work describes how to graphically represent an item that requires the application of multiple abilities as vectors in a multidimensional latent space. The length of the vector for item *i* is equal to the degree of multidimensional discrimination, MDISC. This can be computed using the formula

$$MDISC_i = \sqrt{a_{1i}^2 + a_{2i}^2} \quad (2)$$

MDISC is analogous to the unidimensional IRT model's discrimination parameter. The measurement direction of the vector in degrees from the positive θ axis is

$$\alpha_i = \arccos \frac{a_{1i}}{MDISC_i} \quad (3)$$

This *reference angle* represents the composite of the $\theta - \eta$ ability space that item *i* is best measuring.

The item vector originates at, and is graphed orthogonal to, the p=.5 equiprobability contour. In the compensatory model described in (1) these equiprobability contours are always parallel.

For item *i*, the distance, D_i, from the origin to the p=.5 contour, is computed as

$$D_i = \frac{-d_i}{MDISC} \tag{4}$$

D_i is analogous to the unidimensional IRT difficulty parameter. Because the discrimination parameters can never be negative, the item vectors can lie only in the third quadrant (representing easy items) or in the first quadrant (representing more difficult items). Figure 1 illustrates the item response *surface* for a M2PL item vector whose parameters are: a_1=1.8, a_2=.3, and d=.5. Also illustrated in the bottom portion of Figure 1 is the item's vector, superimposed upon the equiprobability contours of the response surface.

Insert Figure 1 about here

Definition of bias

Bias, according to Shealy & Stout (1989,1991) and Kok (1988), should be conceptualized by examining the difference in certain marginal item characteristic curves (ICCs) for the two groups of interest. The marginal ICC for a particular group is computed by

$$P(X_i = 1 \mid \Theta = \theta) = \int P_i[\theta,\eta]\, f(\eta|\theta)\, d\eta \tag{7}$$

where $P_i(\theta,\eta)$ is the M2PL response function defined in (1) and $f(\eta|\theta)$ is the specified group's conditional distribution of the nuisance dimension, η, given a fixed value of θ, the target ability. For a fixed θ, $P_i[\theta,\eta]$ varies with η. For a fixed value of θ, $P_i(X_i = 1|\Theta = \theta)$ is obtained by averaging $P_i[\theta,\eta]$ over η. Specifically, $P_i(X_i = 1|\Theta = \theta)$ is the unidimensional ICC that will be obtained if differences in the nuisance direction are integrated out. It approximates the ICC that would be obtained via calibration using a unidimensionality based computer program such as BILOG (Mislevy & Bock, 1983) if the test were strictly unidimensional (i.e., if there would be no nuisance dimension η). It is important to note that if $f(\eta|\theta)$ is the same for both groups, bias cannot occur because examinees of equal θ ability will have the same probability of getting the item right.

Bias detection methods

Although there has been a proliferation of methods to detect item bias this paper will focus on only two: the Mantel-Haenzsel (MH) strategy (Holland & Thayer, 1988) and Shealy and Stout's SIB (1991). Both of these procedures are nonparametric and thus require no model calibration. They both have IRT justifications and yet because they do not require IRT calibration, they are computationally non-intensive. To facilitate understanding they will be discussed within the IRT context developed above.

The MH procedure, when placed in a unidimensional IRT framework, examines item bias using the one-parameter Rasch model. In this model all items are assumed to be equal in discrimination and to vary only in difficulty, tenuous assumptions at best. As such, the MH procedure is designed to be primarily sensitive to uniform bias. An item displays uniform bias if the ICCs for two groups of interest differ by only a horizontal translation (i.e., they are "parallel" but not coincident). It is important to note that if the response process is modeled using the 2PL or 3PL IRT models, the ICCs may be non-parallel, causing non-uniform bias. Then the IRT MH theory may not apply (see Zwick (1990) for a more complete discussion). By including the suspect item in the matching criterion it can be shown under the Rasch framework (Holland & Thayer, 1988) that when all of the items, except the suspect item, exhibit no bias, the procedure partials out the effect due to impact in the case of the Rasch model.

The MH-CHISQ has an approximate chi-square distribution with one degree of freedom. However, one cannot tell from this statistic whether a significant value means the item favors the reference or the focal group. To determine the direction of bias one could use the MH estimator $\hat{\alpha}_{MH}$, which represents the average factor by which the odds of a reference group member successfully responding to the studied item exceeds the corresponding odds for a *matched* member of the focal group. A $\hat{\alpha}_{MH}$ value greater than one implies that the reference group outperformed the focal group. Because the odds-ratio scale is not symmetric (i.e., it has a scale of 0 to $+\infty$ with $\alpha = 1$ representing no bias) it is convenient to take the log of $\hat{\alpha}_{MH}$. This new statistic, $\hat{\Delta}_{MH}$, indicates the amount of bias.

In the case of the Rasch model, Δ_{MH} and be expressed as

$$\Delta_{MH} = -2.35(b_f - b_r) \quad (8)$$

where b_f and b_r are the difficulty parameters for the marginal ICCs of the studied item given in (1) above for the focal and reference groups, respectively. The Δ_{MH} index represents the difference in the mean horizontal distance between the marginal ICCs. (Note that when $\Delta_{MH} < 0$, the studied item is biased against the focal group.) The horizontal distance between ICCs is used to assess the amount of bias being represented by the differences in the odds ratio at each score level for the two groups of interest. Some studies (Shealy, 1989; Shealy & Stout, 1991) have reported that the MH chi-square procedure is reasonably robust against inflated Type I error when impact is present for many IRT models as well as robust against loss of power when nonuniform bias is present (even if the generating model is a 2PL or a 3PL IRT model). Impact is defined as the proportion correct differences that occur only on the valid skill. Zwick's (1990) work shows that if the correct model is 2PL or 3PL, the Type I error for MH can be seriously inflated. Recent work by Roussos (1992) found the MH Type I error rate for many such models to be inflated to a larger degree than the SIB procedure.

Shealy and Stout (1991) have a similar theoretical item (and test) bias index called b_{uni} which, in the IRT context, is the average vertical distance between the marginal ICCs of the studied item with respect to θ (the valid subtest ability). This index has a simple empirical interpretation. It is the average difference in probability of a correct response experienced by the two groups for the studied item with impact partialled out. In this sense, b_{uni} is similar conceptually to the Standardization index (Dorans & Kulick, 1986). Computationally b_{uni}, is expressed by

$$b_{uni} = \int_{-\infty}^{+\infty} [T_R(\theta) - T_F(\theta)] f_F(\theta) d\theta \quad (9)$$

where $T_R(\theta)$ and $T_F(\theta)$ are the marginal ICCs of the suspect item for the reference and focal groups respectively, given by (7) and $f_F(\theta)$ is the θ-marginal density of the focal group. Shealy and Stout also have another index, b_{gen}, which is identical to (9) with the exception that the absolute value of the difference between the two marginal ICCs is computed. This index is designed for cases in which non-uniform bias would occur.

An advantage of using SIB, is that the practitioner can weight differences in item performance at each score level by the proportion of examinees in focal group, or the reference group, or both. This feature, which is also present in the Standardization procedure (Dorans & Kulick, 1986) is not present in the MH procedure. In this study the SIB statistic was always weighted by the proportion of focal group examinees represented in the particular observed score category.

One way to express the potential for bias from the Shealy-Stout perspective is by examining the difference between the expected values of the reference and focal group $\eta|\theta$ conditional distributions. If this difference is not equal to zero at every θ, then there is a potential for bias. By examining the expectation of this difference it becomes quite clear which differences in the underlying ability distributions will produce bias. That is, the difference between the expected value of the conditional distributions for a given value of θ can be expressed as

$$\mathcal{E}[\eta_R|\theta] - \mathcal{E}[\eta_F|\theta] = (\mu_{\eta R} - \mu_{\eta F}) + (\rho_R \frac{\sigma_{\eta R}}{\sigma_{\theta R}})(\theta - \mu_{\theta R}) - (\rho_F \frac{\sigma_{\eta F}}{\sigma_{\theta F}})(\theta - \mu_{\theta F}) \qquad (10)$$

It should be noted that the potential for bias as given by (10) only reflects the size of the difference between the conditional expected values of $\eta|\theta$. If the conditional expected values are equal at every θ for the two groups of interest, the potential for bias may still exist because higher order conditional moments need not be the same for the two groups (although is probably unlikely in actual applications). That is, the potential for bias exists whenever the underlying ability distributions for the studied groups are not exactly the same (more accurately (10) should be replaced by $\eta_F|\theta \overset{d}{\neq} \eta_R|\theta$). Using the assumption of bivariate normality, one need only specify the first two moments to identify the underlying ability distributions exactly. Thus, $\eta_F|\theta \overset{d}{=} \eta_R|\theta$ if and only if $\mathcal{E}[\eta_R|\theta] = \mathcal{E}[\eta_F|\theta]$ and $\sigma^2_{\eta_R|\theta} = \sigma^2_{\eta_F|\theta}$.

The Shealy and Stout test statistics and the corresponding estimators, $\hat{b}_{uni}$ and $\hat{b}_{gen}$, are relatively new and do offer the researcher several advantages over the MH procedures and

corresponding estimators $\hat{\alpha}_{MH}$ *and* $\hat{\Delta}_{MH}$. They were developed from a multidimensional modeling perspective and emphasize the examination of bias at the test level.

Method

To study the relationship between level of test reliability as measured by KR-20 (a lower bound estimate) and the MH and SIB statistics, a monte carlo format was used. There were four main factors of interest in the design: amount of bias, number of biased items, number of subjects in the reference and focal groups, and the level of test reliability. These are summarized in Figure 2.

Insert Figure 2 about here

A test with 40 items was simulated for each possible combination of the four factors. All valid items were measuring only θ, the purported skill. Specifically, all items except for the biased item(s) had an a_{2i} parameter equal to zero. Also, for each cell of this fully crossed design the, reference group and focal group examinees were randomly generated from ability distributions which had σ_θ^2 and σ_η^2 equal to 1.5 and .75, respectively. The centroid of the reference group was located at $\mu_\theta = 0.0$, $\mu_\eta = .75$. The focal group was centered at $\mu_\theta = 0.0$, $\mu_\eta = 0.0$. The two latent abilities, θ and η, were correlated $r = .4$ for each group. A density contour for each group is displayed in Figure 3

Insert Figure 3 about here

It should be apparent from the preceding discussion that based upon the marginal distributions for each group, both should perform similarly on any item measuring only θ, but any item capable of measuring η would favor the reference group.

The degree of bias was measured by the extent to which the biased item exploited the difference of the underlying η distributions for the reference and focal groups. This amount ranged from 0° (representing no bias) to 90° (representing the maximum potential for bias) in ten-degree increments.

Valid, non-biased items measured only the first dimension (i.e., $a_2 = 0.0$). For the case in which there were no biased items, one of the valid items was randomly selected to be the suspect item. In the one-biased-item-case the M2PL parameters for the biased item a_1 and a_2 were specifically chosen to keep the MDISC value constant. As such, even though ten different measurement angles were selected for the biased item, the amount of overall discrimination (MDISC) was held constant at 1.5. The one-biased item was always given a d = 0.0 value. For the three-biased-items case, the biased items had the same parameters, identical to the values used in the one-biased-item case.

There were three different levels for the number-of-bias-items factor: 0, 1, and 3. The cases in which no biased items were present were used to examine the Type I error rate of the MH and SIB statistics. Simulations which had one and three biased items were used to examine the power of each statistic. Only the power of SIB was estimated for the cases involving three biased items.

The number-of-subjects factor had six levels (reference n/focal n): 250/250, 500/250, 1000/250, 500/500, 1000/500, 1000/1000. These ratios were selected to cover the approximate size of examinee populations that one might encounter in a national test administration, a state-wide assessment, or a large urban school district setting.

Four levels of reliability were studied: .70, .80, .90, and .95. These four levels of reliability were selected to represent a range of reliabilities one might encounter using different types of tests such as a personality measure or an achievement test. A FORTRAN program was written to randomly select IRT item parameters which would provide the different levels of reliability for the two specified groups. Discrimination parameters were randomly selected from uniform distributions, each having a different range, and difficulty parameters from a N(0,1) distribution. By varying the spread of the a-parameter values, four 40-item tests having the specified KR-20 value were created by trial and error. Because the groups did not differ in their θ abilities, at each reliability level the set of item parameters elicited the same level of reliability

for each group. The effect of replacing one or three of the valid items with a biased item(s) that took advantage of the η-ability differences, resulted in only slight differences ($< .03$) in group reliabilities for the entire 40-item test.

The research design was completely crossed with 720 possible testing conditions. For each testing condition, forty-item tests were randomly simulated for the reference and focal groups and the corresponding MH and SIB statistics were computed. This was replicated 100 times for each condition so that empirical Type I error rates and correct rejection rates could be calculated. A statistical level of significance (α) of .05 was used to test the null hypothesis of no difference in item performance against a non-directional alternative hypothesis.

Results

The means and variances of the M2PL a_{1i} discrimination parameters ($a_{2i} = 0.0$) for each level of reliability are shown in Table 1. These values are computed on the 39 valid items used in the one-biased-item-case. It is interesting to note that as reliability increased the a_{1i} discrimination parameters not only increased but became less variable. In a similar fashion, with each incremental raise in reliability the location of the items (as represented by $\bar{d}_i$) shifted towards the θ mean of the underlying ability distribution and also became less variable.

Insert Table 1 about here

It should be noted that the M2PL model reduces to the unidimensional 2PL IRT model when $a_{2i} = 0.0$. Consequently, to gain more insight into how the items were altered to produce different levels of reliability, plots of the 2PL item characteristic curves (ICCs) were constructed at each level of reliability. These four plots are displayed in Figure 4.

Insert Figure 4 about here

As might be expected by using a reliability of internal consistency, the ICCs become more homogeneous as reliability increases. At the highest levels of reliability, $\rho_{xx'} = .90$ and $\rho_{xx'} = .95$, the ICCs are essentially parallel (Rasch-like) and are located over the center of the underlying ability distribution, indicated by the marginal density curve at the bottom of each plot.

The two-dimensional perspective of the $\rho_{xx'} = .90$ showing the item vectors is illustrated in Figure 5. In this figure a biased item with a 50° orientation is also displayed. Although less dramatic than the unidimensional ICCs in Figure 4, it helps to see the contrast of a test composed of strictly unidimensional items and one biased item. The greater the measurement angle with the positive θ-axis, the less the item contributes to the measurement of θ and the greater its potential for bias. At 90° the biased item is capable of measuring only the nuisance skill. Despite an item's capacity to discriminate between levels of the nuisance ability, bias can only be realized if the two groups of interest differ in their η ability.

Insert Figure 5 about here

A final series of plots were created to examine the change in expected observed score distribution as reliability was increased. These graphs parallel the work by Lord (1953). Shown in Figure 6, each graph displays the test characteristic curve, the (coincident) marginal θ distribution of the focal and reference groups (below the θ axis), and the distribution of the proportion correct true score, ζ (to the left of the ζ axis). These plots are interesting because they graphically demonstrate what happens to the expected raw score distributions as the reliability is increased. Based on the apparent pattern, cases of the very low reliability or extremely high reliability result in expected raw score distributions which have fewer observed score categories to contribute to the computation of the bias detection methods because of the clustering of subjects at particular score levels.

Insert Figure 6 about here

The results of the simulation runs for the no-biased item and one-biased item cases are displayed in Tables 2 and 3. The percent of correct rejections (α = .05) are displayed for each level of reliability for the ten different degree measures of the biased item, for each of the specified reference-focal group sample sizes.

Insert Tables 2-3 about here

The Type I error rate for both statistics appears to be less than the nominal .05 level with only a few exceptions. It does not appear to be influenced by the sample size ratio nor the level of reliability. Likewise, the difference between error rates for SIB and MH does not seem to follow a consistent pattern, nor does there seem to be a significant difference in magnitude.

The pattern for power is clear. As was to be expected, when sample size was reduced the power of both statistics decreased. One way to measure this reduction is to note how large the angle of measurement of the suspect item had to be before the bias procedure was able to determine the item was biased 100% of the time. As the sample size decreased the angle of the biased item at which 100% power was achieved increased for both statistics. For the MH statistic for the 1000/1000 case and $\rho_{xx'}$=.95, 100% rejection rate was obtained for any biased item greater than or equal to 20°. In the case with the smallest sample size (250/250) and $\rho_{xx'}$ = .95 this rate of rejection was not achieved until the suspect item had an angle of 60° or larger. In a similar fashion, SIB achieved a perfect rejection rate for suspect items having an angle of 30° or more with sample sizes of 1000/1000 ($\rho_{xx'}$ = .70, .90 and .95) and 1000/500 ($\rho_{xx'}$ = .70 and .80). For the 250/250 SIB achieved a 100% rejection rate for angles greater than or equal to 60° for $\rho_{xx'}$=.70, .90, and .95.

For any one given sample size and angular direction of the biased item, the power of each statistic increased as the reliability increased, although not in a consistent manner. In the 1000/250 case with the biased item at 20° there was a drop at higher levels of reliability; the MH power rates were .66, .71, .68, and .66 for $\rho_{xx'}$ values of .70, .80, .90, and .95,

respectively. For these same conditions the SIB power rates were .70, .72, .69, and .72. This drop in power occurs again for both statistics at the smallest simulated sample size, 250/250.

Differences between the MH and SIB procedures in the rate of correct rejections always seem to be quite small and not consistently in one direction. The largest differences occur in the 250/250 case with $\rho_{xx'} = .90$ and the biased item having angles of 20° (MH = .46, SIB = .56) and 30° (MH = .86 and SIB = .79).

The results for the cases in which three items were biased are shown in Tables 4 and 5. Only the power of SIB was evaluated in these cases. The Type I error rate was less than the nominal .05 level with only a few exceptions, namely the 1000/250 case with $\rho_{xx'} = .90$ for which the Type I error rate was .10. Averaged over all conditions the Type I error rate was about .04.

Insert Tables 4 and 5 about here

For comparable sample sizes and levels of reliability the rejection rates of SIB in the three-biased-item cases are much higher than the one-biased item cases. At each level of reliability, SIB achieves a 100% rejection rate when the three items have an angle of 20° or greater for sample sizes of 1000/1000, 1000/500, 500/500, and 500/250. For the smallest sample size, this rate was achieved for angles greater than or equal to 40°. In all cases as reliability increased power increased.

Discussion

The purpose of this study is to provide the practitioner with a set of guidelines concerning the power of the MH and SIB tests for various levels of reliability and sample sizes. It is intended to highlight conditions for which the MH and SIB statistics may yield inaccurate results. Based upon the results its appears that both procedures are about equally powerful, with the exception at 10° and 20°, where SIB appears to be slightly more powerful. The Type I error

rates for each procedure seem to be below the nominal .05 level, and comparable at all sample size levels.

It appears that for small sample sizes there may be an range of reliability in which power is optimal. Ideally one would want to gather information at all levels of the raw score scale, but as seen in Figure 6, this may not be possible with very low or very high levels of reliability.

In the multiple-biased item case SIB seems to perform quite well. Because this is the only known procedure which can examine multiple items at one time this is encouraging news. Practitioners should be encouraged to examine the effect of several biased items in concert, and SIB seems to be a good procedure for doing this.

One also gains a sense of how abberant angle-wise an item would have to be before it would be consistently rejected as being biased. More work needs to be done to determine the relationship between the angular difference and the substantive meaning of the item. In the real-world tests are not strictly unidimensional and thus an angular difference of 30° may not be considered large enough to make the item construct invalid.

In conclusion, it is important to note that there are a numbers of factors that were not investigated in this study. First, in a real testing situation all test items never have vectors that lie in the same direction. Ideally they will lie in a narrow sector (cf. Ackerman, 1992). How the width of this sector effects the power of each statistic is unknown. Second, it can also be assumed that non-uniform bias does exist in many testing situations. How this effects the correct rejection rates of the MH and SIB procedures at different levels of reliability and for biased items having different measurement angles also remains unknown. Third, one might choose to simulate increasing levels of reliability in a different manner. Specifically, by creating longer tests. Simulations done in this manner may not necessarily imitate reality because of the large number of items needed to achieve high levels of reliability. However, such a method would not have as severe an effect on the observed score distribution as the manipulation of discrimination parameters which was done in this study. Finally, this study looked only at cases in which the test was equally reliable for both groups of interest. Conceivably this may not always be the case.

Clearly bias research has just "scratched the surface" when it comes to understanding the capability of various procedures to successfully determine when an item is biased. One word of caution: in trying to detect biased items, one should never become to involved with the statistics and forget about the actual item. Practitioners should never lose sight of all the factors that could influence the examinees' response patterns (e.g., the wording of an item, its format, its position, etc.).

Table 1
Means and standard deviations of M2PL discrimination parameters for each level of reliability.

KR-20 Reliability	$\bar{a}_1$	$\sigma^2_{a_1}$	$\bar{d}_i$	σ^2_d
.70	.68	.28	.45	1.71
.80	.80	.31	.17	1.43
.90	1.17	.14	-.19	.79
.95	1.71	.10	-.15	.62

Note. n = 40 for each test. a_{2i} = 0.0 for all items.

Table 2

Empirical Type I error and power values for MH and SIB for samples sizes of 1000/1000, 1000/500 and 1000/250 and the case of 1 biased item.

Sample size (Ref/Foc)	Angle[a] of Biased Item	Level of Reliability .70 MH	.70 SIB	.80 MH	.80 SIB	.90 MH	.90 SIB	.95 MH	.95 SIB
1000/1000	00[b]	04	03	03	04	06	05	01	05
	10	53	59	54	58	56	57	56	62
	20	98	98	99	99	99	*	*	*
	30	*	*	*	*	*	*	*	*
	40	*	*	*	*	*	*	*	*
	50	*	*	*	*	*	*	*	*
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*
1000/500	00[b]	02	03	07	05	02	03	03	05
	10	37	40	38	44	37	33	41	44
	20	95	97	93	92	91	92	91	92
	30	*	*	*	*	99	99	99	99
	40	*	*	*	*	*	*	*	*
	50	*	*	*	*	*	*	*	*
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*
1000/250	00[b]	03	06	04	06	03	05	06	05
	10	23	31	28	32	28	27	29	26
	20	66	70	71	72	68	69	66	72
	30	90	94	93	96	97	99	96	97
	40	*	96	*	*	*	*	*	*
	50	*	*	*	*	*	*	*	*
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*

Note. Decimals are deleted. * denotes a value of 1.0.

[a]Angles are expressed in degrees from the positive θ-axis

[b]Denotes the no bias case; corresponding row represents Type I error rate

Table 3

Empirical Type I error and power values for MH and SIB for samples sizes of 500/1000, 500/250 and 250/250 and the case of 1 biased item.

Sample size (Ref/Foc)	Angle[a] of Biased Item	Level of Reliability .70 MH	.70 SIB	.80 MH	.80 SIB	.90 MH	.90 SIB	.95 MH	.95 SIB
500/500	00[b]	01	03	04	06	02	05	02	03
	10	20	25	28	30	24	25	29	39
	20	77	78	79	82	84	84	83	82
	30	98	99	98	97	99	99	99	98
	40	*	*	*	*	*	*	*	*
	50	*	*	*	*	*	*	*	*
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*
500/250	00[b]	01	03	00	01	02	03	01	02
	10	13	19	22	21	17	17	18	17
	20	59	67	62	66	59	64	62	63
	30	93	94	96	93	90	90	92	92
	40	99	99	*	*	*	98	98	99
	50	*	*	*	*	*	99	*	*
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*
250/250	00[b]	04	04	03	06	03	01	01	03
	10	13	22	16	19	12	22	10	18
	20	49	49	54	56	46	56	50	49
	30	84	92	90	87	86	79	82	79
	40	99	96	98	98	98	94	98	98
	50.	*	99	*	*	*	99	99	94
	60	*	*	*	*	*	*	*	*
	70	*	*	*	*	*	*	*	*
	80	*	*	*	*	*	*	*	*
	90	*	*	*	*	*	*	*	*

Note. Decimals are deleted. * denotes a value of 1.0.

[a]Angles are expressed in degrees from the positive θ-axis

[b]Denotes the no bias case; corresponding row represents Type I error rate

Table 4
Empirical Type I error and power values SIB for sample sizes of 1000/1000, 1000/500 and 1000/250 and the case of 3 biased items.

Sample size (Ref/Foc)	Angle[a] of Biased Item	Level of Reliability .70	.80	.90	.95
1000/1000	00[b]	04	06	05	04
	10	89	93	96	95
	20	*	*	*	*
	30	*	*	*	*
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*
1000/500	00[b]	02	04	03	02
	10	79	84	83	89
	20	*	*	*	*
	30	*	*	*	*
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*
1000/250	00[b]	03	06	03	05
	10	51	58	59	60
	20	96	98	98	98
	30	*	*	*	*
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*

Note. Decimals are deleted. * denotes a value of 1.0.
[a]Angles are expressed in degrees from the positive θ-axis
[b]Denotes the no bias case; corresponding row represents Type I error rate

Table 5
Empirical Type I error and power values SIB for sample sizes of 500/500, 500/250 and 250/250 and the case of 3 biased items.

Sample size (Ref/Foc)	Angle[a] of Biased Item	Level of Reliability .70	.80	.90	.95
500/500	00[b]	05	07	07	01
	10	59	66	67	74
	20	*	98	*	*
	30	*	*	*	*
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*
500/250	00[b]	03	10	03	02
	10	47	50	48	51
	20	94	96	95	98
	30	*	*	*	*
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*
250/250	00[b]	04	03	03	04
	10	36	37	44	36
	20	88	92	89	84
	30	99	99	98	99
	40	*	*	*	*
	50	*	*	*	*
	60	*	*	*	*
	70	*	*	*	*
	80	*	*	*	*
	90	*	*	*	*

Note. Decimals are deleted. * denotes a value of 1.0.
[a]Angles are expressed in degrees from the positive θ-axis
[b]Denotes the no bias case; corresponding row represents Type I error rate

References

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. Journal of Educational Measurement 24, 67-91.

Dorans, N.J. & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement 23, 355-368.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunter, J.F. (1975, December). A critical analysis of the use of item means and item-test correlations to determine the presence or absence of content bias in achievement test items. A paper presented at the National Institute of Education Conference on Test Bias. Annapolis, MD.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine and J. Rost (Eds.), Latent trait and latent class models. (pp. 263-274). New York, NY: Plenum Press.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. Educational and Psychological Measurement. 547-549.

Mislevy, R.J. & Bock, R.D.(1983). BILOG: Item analysis and test scoring with binary logistic models [Computer Program]. Mooresville, IN: Scientific Software.

Pine, S.M. (1977). Applications of item response theory to the problem of test bias. In D.J. Weiss (Ed.), Applications of computerized adaptive testing (Research Report 77-1). Minneapolis, MN: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Reckase, M.D. (1985, April). The difficulty of test items that measure more than one ability. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Reckase, M.D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Shealy, R. & Stout, W. (in press). An item response theory model for test bias. (ONR Technical Report); In H. Wainer & P. Holland (Eds.), Differential Item Functioning, Theory and Practice, Hillsdale, NJ: L. Erlbaum Associates.

Shealy, R. & Stout, W. (1991). A procedure to detect test bias present simultaneously in several items. (Technical Report 91-3-ONR). Champaign, IL: University of Illinois.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia, 57-70.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide. Journal of Educational Measurement, 3, 185-197.

Footnotes

[1]Both procedures can be used to detect either bias or DIF. It should be noted that bias and DIF are distinct concepts; see Shealy and Stout (1991) for a careful discussion of the difference. In this paper the term item bias will refers to situations where the user wishes to detect either bias or DIF.

Figure Captions

Figure 1. The item response surface and corresponding contour with the item vector for the M2PL parameters, $a_1=1.8$, $a_2=.3$, $d=.5$.

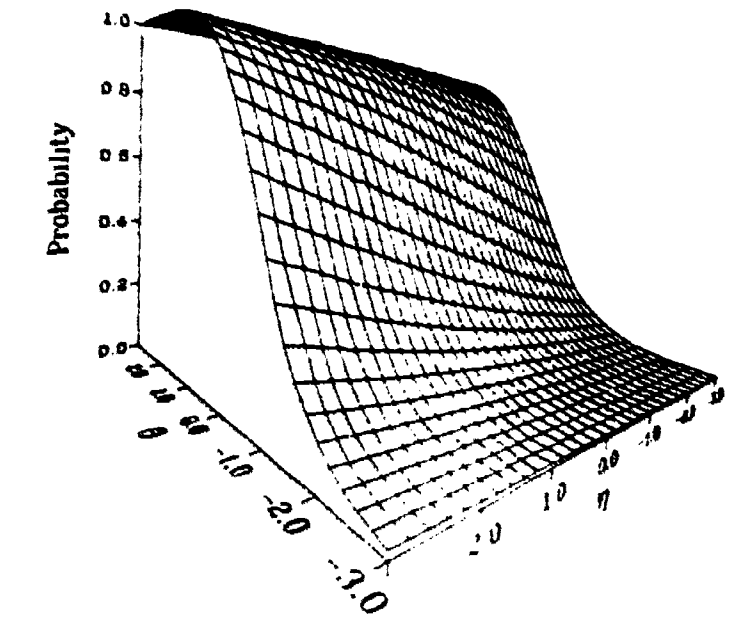
Figure 2. Research design of the study detailing the four factors of interest which were fully crossed.

Figure 3. A contour plot of the densities for the Reference and Focal groups with accompanying marginal distributions.

Figure 4. Unidimensional ICCs for the 39 valid items for tests having KR-20 reliabilities of .70, .80, .90, and .95.

Figure 5. A plot of the 40 item vectors for a KR-20 Reliability of .90 with one biased item having an effect size of 50°.

Figure 6. A plot showing the distributional relationships between the generating θ distribution and the proportion-correct true score distribution.


Probability
1.0
0.8
0.6
0.4
0.2
0.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
θ
η

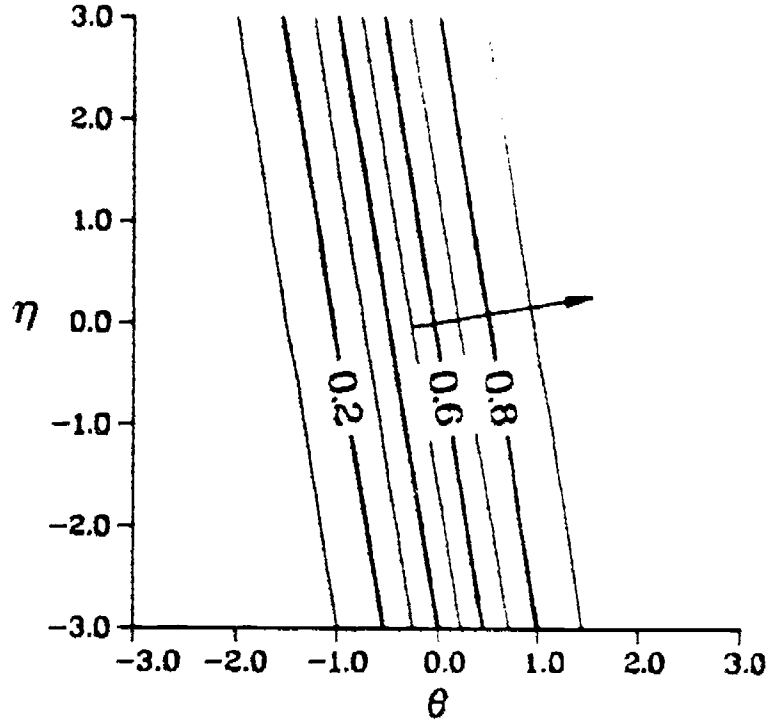
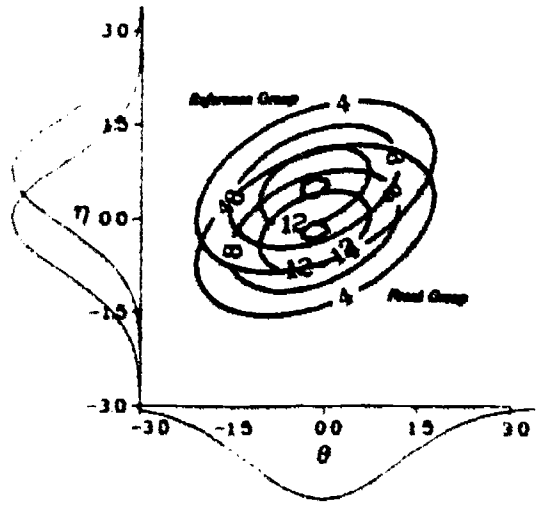

3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
η
−3.0 −2.0 −1.0 0.0 1.0 2.0 3.0
θ
0.2
0.6
0.8

Figure 1

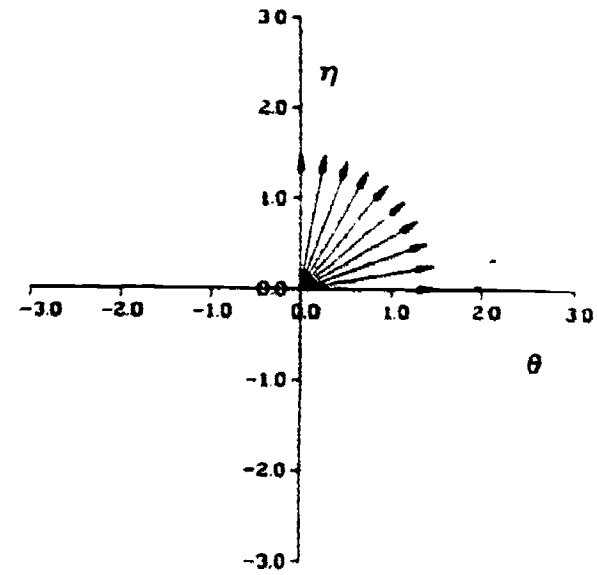
Figure 2

Factor 1: *Sample size - 6 levels* (Ref N/Foc N)



1000/1000
1000/500
1000/250
500/500
500/250
250/250

Factor 2: *Direction of biased item - 10 levels (amount of bias)*



0° → 90° *in* 10° *increments*

Factor 3: *Level of Reliability - 4 levels*

KR-20

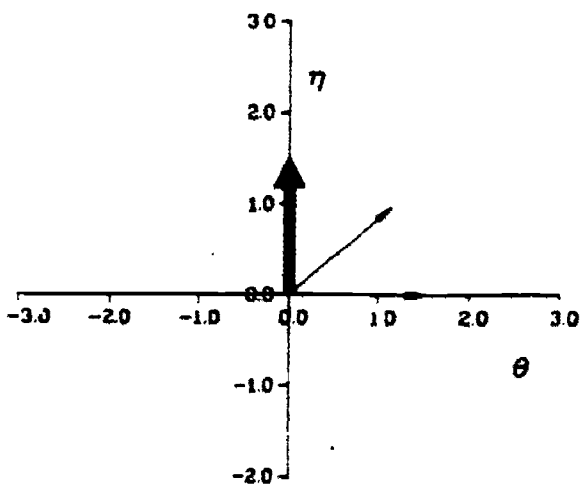
$_{20}\rho_{xx'}$ - .70

$_{20}\rho_{xx'}$ - .80

$_{20}\rho_{xx'}$ - .90

$_{20}\rho_{xx'}$ - .95

Factor 4: *Number of biased items - 3 levels*



0 (To evaluate Type I error rates)

1 (To evaluate Power of MH & SIB)

3 (To evaluate Power of SIB)

Figure 3

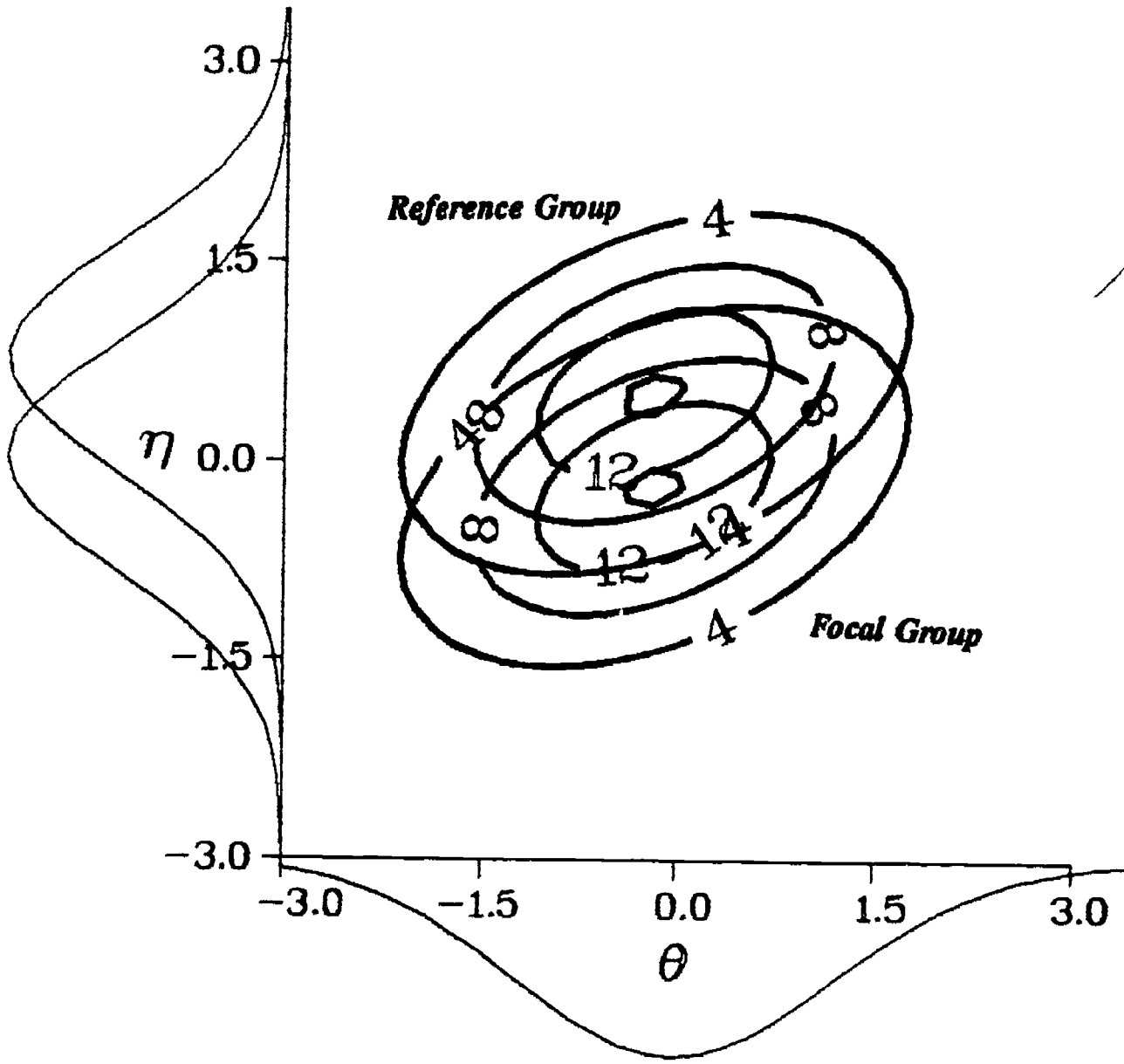
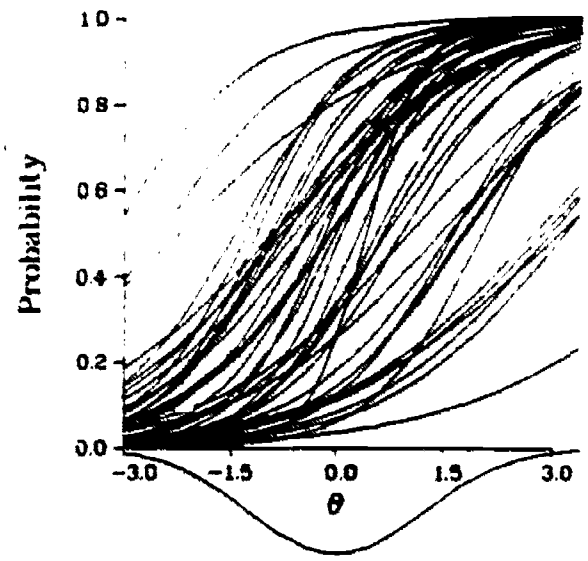
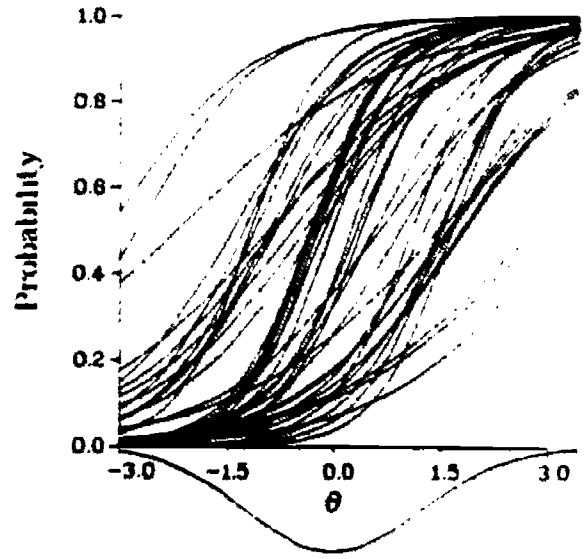

Reference Group
Focal Group
3.0
1.5
0.0
−1.5
−3.0
η
−3.0
−1.5
0.0
1.5
3.0
θ
4
8
12
14

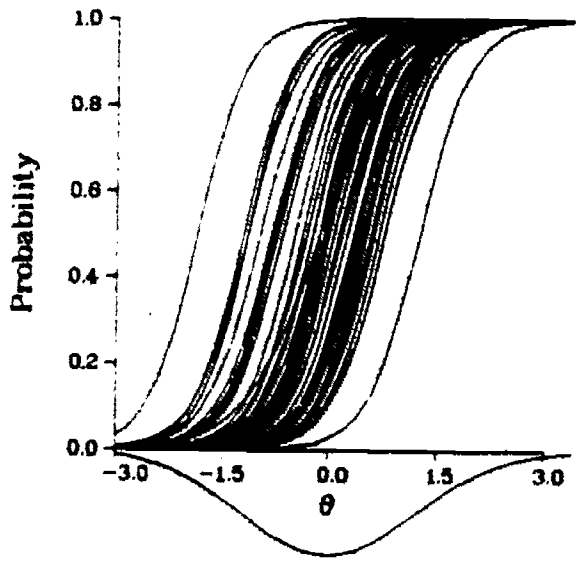
Figure 4



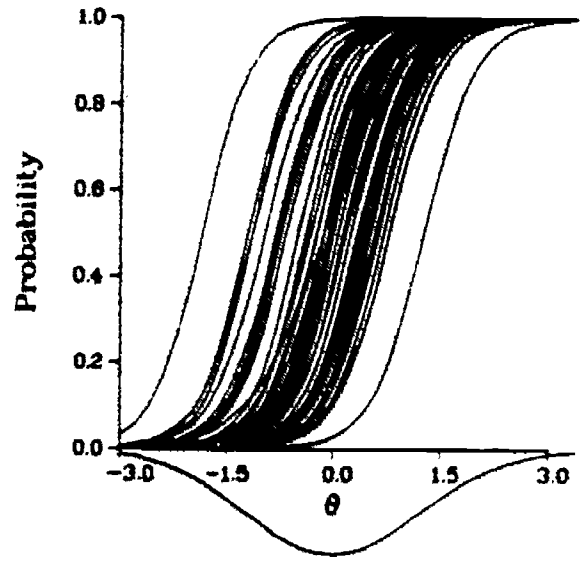
$2\rho_{xx'}$ = .70



$2\rho_{xx'}$ = .80

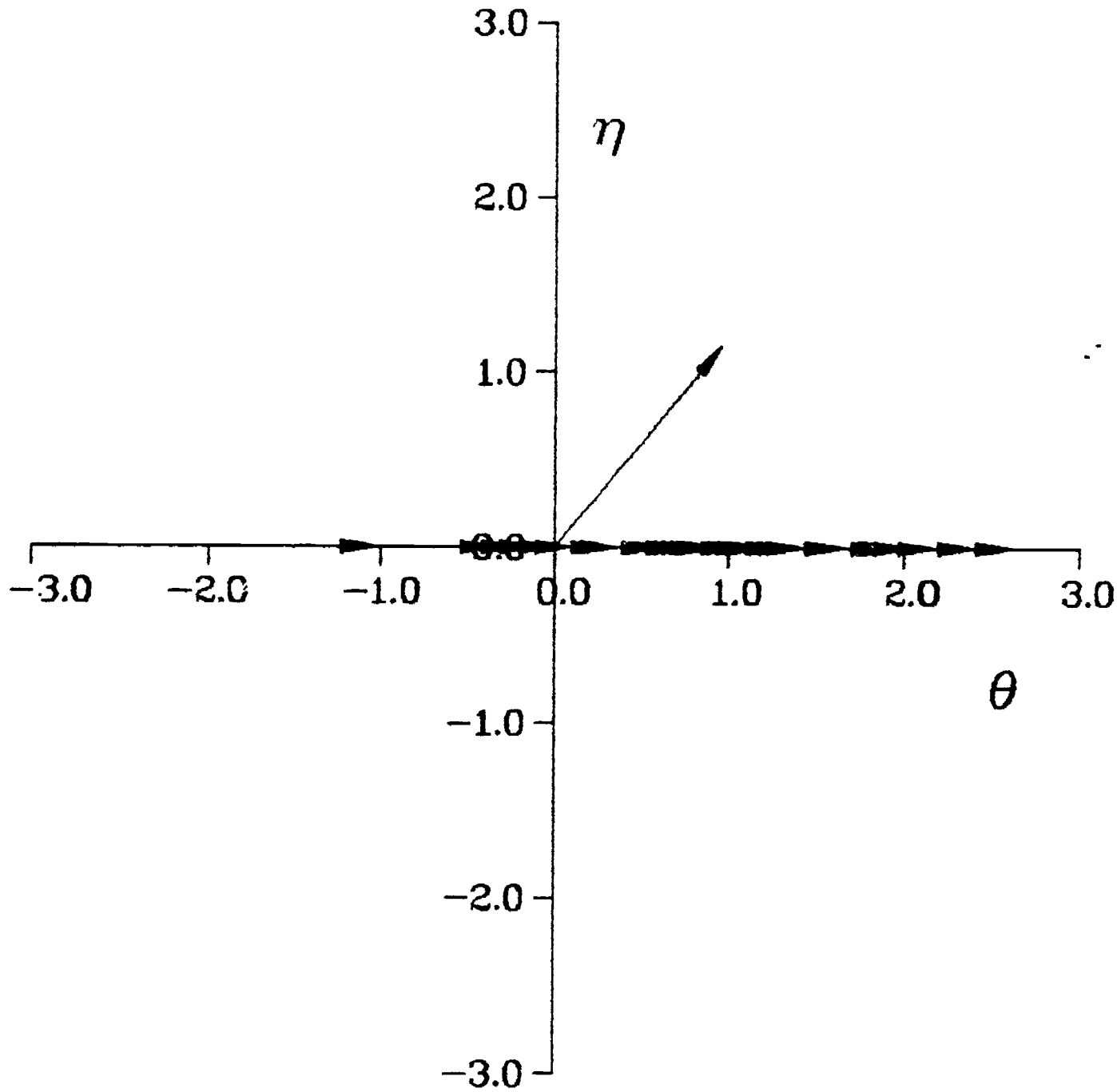


$2\rho_{xx'}$ = .90

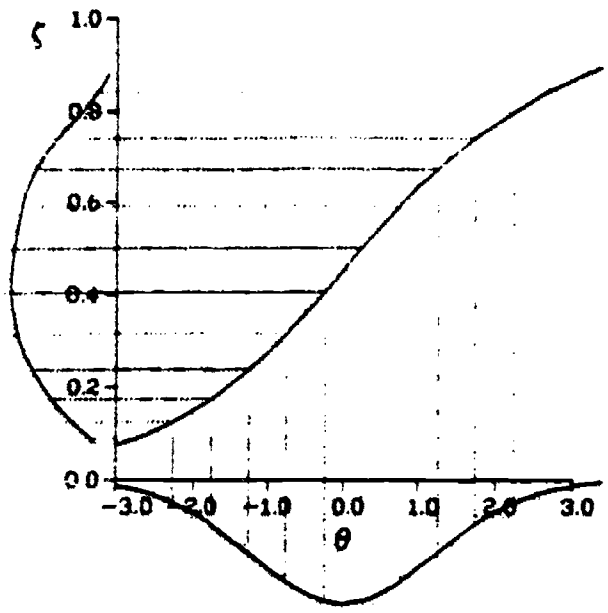


$2\rho_{xx'}$ = .95

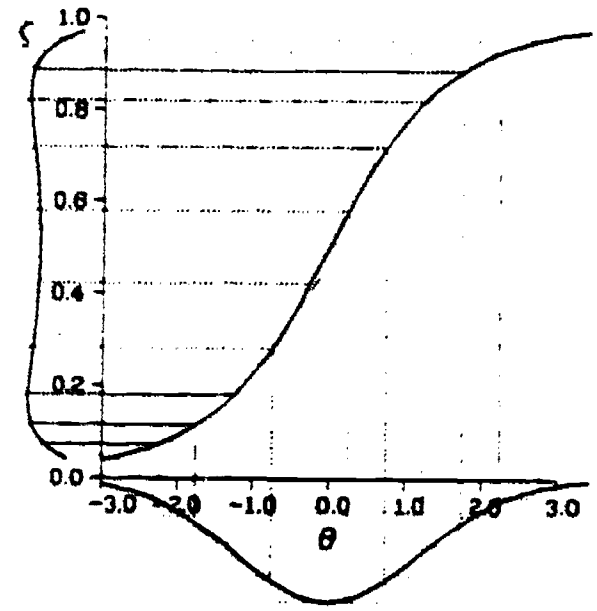
Figure 5



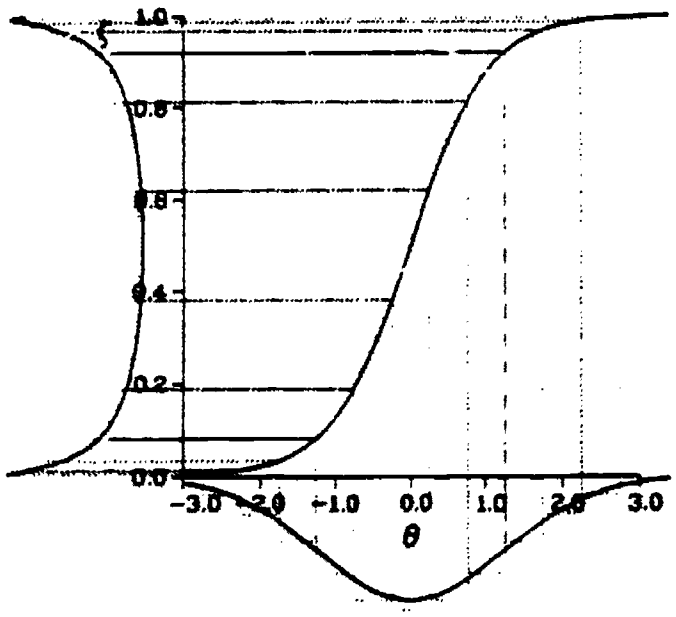
$\rho_{xx'} = .70$



$\rho_{xx'} = .80$



$\rho_{xx'} = .90$



$\rho_{xx'} = .95$

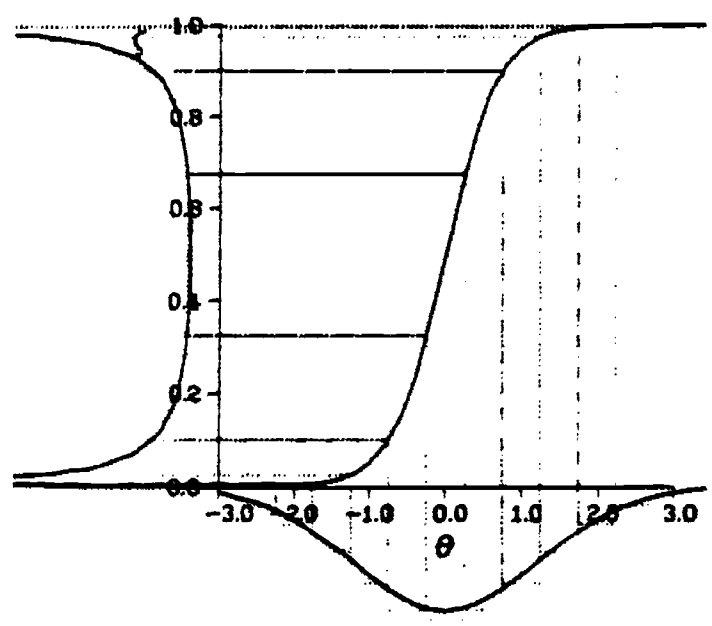


Figure 6